Computational Statistics and Data Analysis 56 (2012) 559–573

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Nonparametric regression models for right-censored data using Bernstein polynomials

Muhtarjan Osman *, Sujit K. Ghosh

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

ARTICLE INFO

Article history:
Received 8 February 2011
Received in revised form 29 June 2011
Accepted 31 August 2011
Available online 18 September 2011

Keywords:
Bernstein polynomials
Censored data
Nonparametric regression
Nonproportional hazards
Sieve

ABSTRACT

In some applications of survival analysis with covariates, the commonly used semiparametric assumptions (e.g., proportional hazards) may turn out to be stringent and unrealistic, particularly when there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. We present a new nonparametric regression model for the conditional hazard rate using a suitable sieve of Bernstein polynomials. The proposed nonparametric methodology has three key features: (i) the smooth estimator of the conditional hazard rate is shown to be a unique solution of a strictly convex optimization problem for a wide range of applications; making it computationally attractive, (ii) the model is shown to encompass a proportional hazards structure, and (iii) large sample properties including consistency and convergence rates are established under a set of mild regularity conditions. Empirical results based on several simulated data scenarios indicate that the proposed model has reasonably robust performance compared to other semiparametric models particularly when such semiparametric modeling assumptions are violated. The proposed method is further illustrated on the gastric cancer data and the Veterans Administration lung cancer data.

1. Introduction

In regression analysis of survival data, the proportional hazard model (Cox, 1972) has become by far the most widely used method by researchers in many disciplines especially in the field of biostatistics. The most appealing features of the Cox model include the unspecified baseline hazard function and the straightforward interpretation for the effect of categorical covariates such as treatment assignment. In many applications the proportional hazard (PH hereafter) assumption is generally considered reasonable for the cases with a long follow-up time (Perperoglou et al., 2007). In some other situations, however, the validity of this simplification is certainly questionable. It is known that the violation of the PH assumption could lead to erroneous inference in some circumstances (see e.g., Schemper, 1992). In these cases, several alternatives such as proportional odds (PO hereafter) model (Bennett, 1983) and accelerated failure time (AFT hereafter) model (Kalbfleisch and Prentice, 1980) have been proposed. But these semiparametric models may also turn out to be stringent or even unrealistic in some cases. For instance, consider the case of crossing survival curves, in which there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. For example, in the well known gastric cancer clinical trial (Stablein et al., 1981), patients receiving only chemotherapy may have higher survival rates initially but such rates decay much faster compared to the group of patients receiving chemotherapy and radiotherapy. None of the above models (PH, PO, AFT) can accommodate such a feature of the data. The cases with nonproportional hazards can also occur in non-medical applications. For example, Dolton and O'Neill (1996) reported a non-proportional effect of government official counseling on the hazard function of unemployment benefits sign-offs in the United Kingdom.

* Corresponding author.
E-mail addresses: mosman@ncsu.edu (M. Osman), sujit_ghosh@ncsu.edu (S.K. Ghosh).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.08.019

In the Cox PH model, the conditional hazard function $h(t|Z)$ is modeled as $h(t|Z) = h_0(t)\exp(\beta^T Z)$ where $h_0(\cdot)$ denotes the baseline hazard, Z represents the vector of covariates and the parameter of interest β is constant over time. Several semiparametric extensions of the Cox PH model have been proposed by various authors to relax the proportionality assumption. The most popular approach is to include a time-varying effect $\beta(t)$ by replacing β in the Cox PH model. A challenging step in the time-varying coefficient model is the estimation of the effect function $\beta(\cdot)$. Murphy and Sen (1991) assumed $\beta(\cdot)$ is piecewise constant and proposed a histogram sieve estimator. Zucker and Karr (1990) used smoothing splines based on partial likelihood. The time-varying coefficient model has received extensive attention in literature recently. For further details on more recent approaches on this subject, we refer to the papers by Martinussen and Scheike (2002), Cai and Sun (2003) and Tian et al. (2005). Most of the methodologies involving the time-varying effect may turn out to be computationally intensive because the form of likelihood function is usually very complicated. Besides, Perperoglou et al. (2007) pointed out that when survival curves cross, over-emphasizing on the regression coefficient might not be appropriate. The reason is that the appealing feature of easy interpretation and estimation of the Cox PH model, which comes from the separation of time and covariate effects, will be lost under nonproportional hazards. Instead, the entire conditional hazard or survival curve will be more informative to medical researchers. Therefore, in this paper we take a different direction by directly modeling the conditional hazard function using Bernstein polynomials.

In contrast to various extensions of the Cox PH model to accommodate nonproportionality, our method is completely nonparametric and computationally much simpler to implement. Although nonparametric models suffer from the curse of dimensionality, as Spierdijk (2008) pointed out they can serve as starting points for building parametric or semiparametric models in high dimensions. Fully nonparametric hazard regression models have been studied by many authors, most of which focused on the kernel method and smoothing splines (see e.g., Li and Doss, 1995; Gray, 1996; Spierdijk, 2008, among others). As pointed out by Tenbusch (1994, 1997), the estimators based on Bernstein polynomials can be regarded as kernel-based estimators with spatial adaptive polynomials as weight or kernel functions. So instead of using the same kernel function, the kernel function in the Bernstein polynomial based estimator adapts itself to the positions of the knots specified. It is also this property makes the estimators based on Bernstein polynomials enjoy better boundary behavior than the usual kernel-based estimators. Another important methodology in this line of research is referred to as HARE (Hazard Regression) in Kooperberg et al. (1995). HARE is a regression model based on linear splines and their tensor products for the conditional log-hazard function. In the HARE model, linear splines are used rather than quadratic or cubic splines in order to avoid numerical integration in the log-likelihood (and in its gradient and Hessian matrix). This simplification is essentially due to the model selection step involving stepwise addition and stepwise deletion incorporated in HARE. However, in some situations where the conditional log-hazard function takes complex form the linear splines and their tensor products may not capture the overall dependence of the event time on other covariates.

Bernstein polynomials have been considered in a wide range of statistical problems based on completely observed data. The most common application is density estimation, which dates back to the work of Vitale (1975). Some of the most recent work on this topic includes Petrone (1999), Babu et al. (2002) and Choudhuri et al. (2004) among many others. Bernstein polynomials have also been applied in the regression setting by Tenbusch (1997) and Chang et al. (2007). In the context of survival analysis with censored data, Chang et al. (2005) proposed using Bernstein polynomials for hazard rate estimation in a Bayesian framework for a homogeneous population, i.e., without any covariates.

In this paper, we consider nonparametric hazard regression based on Bernstein polynomials for right-censored data. As we will demonstrate later, Bernstein polynomials have several advantages in this particular setting. Monotonicity of the cumulative hazard function can be modeled naturally via Bernstein polynomials. In addition, Bernstein polynomials have nice differentiability properties such that the log-likelihood, its gradient, and Hessian matrix all take relatively easy forms, making our method very easy to implement as compared to other computationally intensive methods such as those based on the time-varying coefficient models. To obtain a smooth estimator for the conditional hazard function in a full nonparametric setting, we use a sieve maximum likelihood estimator (Grenander, 1981; Geman and Hwang, 1982). The proposed nonparametric regression model in this paper is shown to encompass a proportional hazards structure. The rest of the paper proceeds as follows. In Section 2, we describe the model for categorical and continuous covariates. We show that the sieve maximum likelihood estimate is consistent and the corresponding rate of convergence is derived in Section 3. In Section 4, the proposed method is demonstrated through simulated data as well as real data from two well known cancer studies. Finally, we conclude with some discussions in Section 5.

2. Conditional hazard model using Bernstein polynomials

First, we consider the one-sample right-censored data with no covariates. Suppose an experiment or a clinical trial consists of n subjects, T_i denotes the continuously distributed time to certain event of interest for the subject i where $i = 1, 2, \ldots, n$. The event time T_i is subject to random right-censoring C_i and hence for each subject we observe (X_i, Δ_i), where $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and $I(A)$ denotes the indicator function that takes the value 1 when the event A is true, otherwise $I(A) = 0$. We also assume that T_i is statistically independent of C_i for each $i = 1, 2, \ldots, n$. For any $t \geq 0$, the cumulative hazard function is given by $H(t) = -\log S(t)$ and the hazard function $h(t) = \dot{H}(t)$, where $S(t) = \Pr(T_i > t)$ is the survival function and $\dot{H}(t)$ denotes the derivative of $H(t)$. Further, following standard practice (e.g. Tian et al., 2005) we assume that there exists a $\tau < \infty$ such that $\tau = \inf\{t : S(t) = 0\}$.